PLOS ONE

RESEARCH ARTICLE

Linkage disequilibrium vs. pedigree: Genomic selection prediction accuracy in conifer species

Frances R. Thistlethwaite[1], Omnia Gamal El-Dien[1,2], Blaise Ratcliffe[1], Jaroslav Klápště[3], Ilga Porth[4], Charles Chen[5], Michael U. Stoehr[6], Pär K. Ingvarsson[7], Yousry A. El-Kassaby[1]*

1 Department of Forest and Conservation Sciences, Faculty of Forestry, The University of British Columbia, Vancouver, British Columbia, Canada, **2** Pharmacognosy Department, Faculty of Pharmacy, Alexandria University, Alexandria, Egypt, **3** Scion (New Zealand Forest Research Institute Ltd.), Whakarewarewa, Rotorua, New Zealand, **4** Département des sciences du bois et de la forêt, Université Laval, Québec, QC, Canada, **5** Department of Biochemistry and Molecular Biology, Oklahoma State University, Stillwater, OK, United States of America, **6** British Columbia Ministry of Forests, Lands and Natural Resource Operations, Victoria, BC, Canada, **7** Linnean Centre for Plant Biology, Department of Plant Biology, Swedish University of Agricultural Sciences, Uppsala, Sweden

* y.el-kassaby@ubc.ca

OPEN ACCESS

Citation: Thistlethwaite FR, Gamal El-Dien O, Ratcliffe B, Klápště J, Porth I, Chen C, et al. (2020) Linkage disequilibrium vs. pedigree: Genomic selection prediction accuracy in conifer species. PLoS ONE 15(6): e0232201. https://doi.org/10.1371/journal.pone.0232201

Editor: Dusan Gomory, Technical University in Zvolen, SLOVAKIA

Received: January 29, 2020

Accepted: April 8, 2020

Published: June 10, 2020

Data Availability Statement: All relevant data are in Dryad: Douglas-fir: (https://doi.org/10.5061/dryad.8n2d374) Interior spruce: (http://doi.org/10.5061/dryad.8kb37).

Funding: This work was supported by Genome British Columbia (User Partnership Program (UPP-001) to YAK and MUS, NSERC Discovery Grant to YAK. We declare that the funding agencies did not participate in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Abstract

Background

The presupposition of genomic selection (GS) is that predictive accuracies should be based on population-wide linkage disequilibrium (LD). However, in species with large, highly complex genomes the limitation of marker density may preclude the ability to resolve LD accurately enough for GS. Here we investigate such an effect in two conifer species with ~ 20 Gbp genomes, Douglas-fir (*Pseudotsuga menziesii* Mirb. (Franco)) and Interior spruce (*Picea glauca* (Moench) Voss x *Picea engelmannii* Parry ex Engelm.). Random sampling of markers was performed to obtain SNP sets with totals in the range of 200–50,000, this was replicated 10 times. Ridge Regression Best Linear Unbiased Predictor (RR-BLUP) was deployed as the GS method to test these SNP sets, and 10-fold cross-validation was performed on 1,321 Douglas-fir trees, representing 37 full-sib F_1 families and on 1,126 Interior spruce trees, representing 25 open-pollinated (half-sib) families. Both trials are located on 3 sites in British Columbia, Canada.

Results

As marker number increased, so did GS predictive accuracy for both conifer species. However, a plateau in the gain of accuracy became apparent around 10,000–15,000 markers for both Douglas-fir and Interior spruce. Despite random marker selection, little variation in predictive accuracy was observed across replications. On average, Douglas-fir prediction accuracies were higher than those of Interior spruce, reflecting the difference between full- and half-sib families for Douglas-fir and Interior spruce populations, respectively, as well as their respective effective population size.

Competing interests: No competing interests.

Conclusions

Although possibly advantageous within an advanced breeding population, reducing marker density cannot be recommended for carrying out GS in conifers. Significant LD between markers and putative causal variants was not detected using 50,000 SNPS, and GS was enabled only through the tracking of relatedness in the populations studied. Dramatically increasing marker density would enable said markers to better track LD with causal variants in these large, genetically diverse genomes; as well as providing a model that could be used across populations, breeding programs, and traits.

Introduction

With genotyping costs at the lowest they have ever been (and still on a decreasing trajectory), genomic selection (GS) becomes an ever-more viable option for forest tree breeders. This should subsequently result in beneficial gains to the industry in terms of improved wood quality, yield per unit time (generational turnover), and stress tolerance (biotic and abiotic) [1,2]. In a deviation from marker-assisted selection (MAS) [3], rather than attempting to identify significant trait-loci relationships, GS employs all available marker information simultaneously to predict traits performance [4]. GS experimental results have been promising to date, with prediction accuracies higher than those of MAS [5].

One of the major determinants of GS success is the relationship between effective population size (N_e) and marker density [6]. Falconer and Mackay [7] succinctly describe N_e as the number of randomly mating individuals that would cause the observed inbreeding rate of a population. As a result of inbreeding, which is more likely at low N_e, allele frequencies become skewed. In this situation (at low N_e) certain individuals have an increased chance to reproduce, causing genetic drift as their genes are passed more frequently onto the next generation [6]. Over time and over multiple generations of these conditions those alleles may become fixed, as genetic diversity decreases [7]. Low N_e populations are subject to stronger drift, which in turn is one of the major driving forces influencing linkage disequilibrium (LD) between loci. Lower N_e populations have higher LD, and this LD between markers and trait QTLs is essential for the prediction of trait performance from markers [6]. The success of GS is highly dependent on the marker-QTL LD and this is largely determined by the extent to which the training and validation populations are related. A certain amount of caution is also required during the implementation of GS as LD will decay subsequent to every round of breeding, due to recombination, although this can be overcome by employing dense marker arrays. It should be emphasized that GS may require alternative delivery methods for tree improvement, as the current seed orchard production pathway and its sexual-production mode effectively breaks down the marker-QTL LD. Lin et al. [6] point out that N_e can be artificially reduced by using half-sib or within-family populations. This is a good technique to use on outbred species, such as forest trees, when applying GS.

The determination of the number of markers required for GS is largely based on the occurrence of LD, which in turn is determined by population structure and N_e. Solberg et al. [8] did some initial investigations into marker density effect, comparing two types of marker: SSR-like multiallelic markers versus SNP-like biallelic markers. They found a general increase in prediction accuracy as marker density increased in relation to N_e, however they failed to reach a plateau with the numbers available ($2N_e$ SSR markers per Morgan or $8N_e$ SNP markers per Morgan). Meuwissen [9] proposed that a minimum of $10N_eL$ markers should be used to obtain

accurate predictions in GS, where L is the total length of the genome in Morgans. As Lin et al. [6] discuss, this should mean that for a population of outbreeding apple (*Malus* sp.) trees with an assumed N_e of 1,000 and a genome of approximately 13 Morgans, 130,000 markers are necessary for accurate GS predictions. However Kumar et al. [10] successfully carried out GS in apple variety *Malus domestica* Borkh with an accuracy of 0.7, using only 2,500 markers. This can be attributed to the bi-parental design used [9]. Within-family designs call for fewer markers since larger (but fewer) chromosomal segments are shared by family members, and it is these that need to be tracked [11]. Grattapaglia and Resende [12], using a deterministic approach found that for GS to be effective, a marker density of ~2 markers/cM is required when N_e is no greater than 30 and larger N_e may require up to 20 markers/cM. Yet higher marker densities may be required for situations in which the training and validation populations are not derived genetically from the same base population [9]. Using Grattapaglia and Resende's [12] calculations, and an assumed map length of 2,000 cM [13–16], we should aim to use 4,000 markers for investigations into Douglas-fir (*Pseudotsuga menziesii* Mirb. (Franco)) with an N_e ~21. By the same reasoning we should aim to use up to 40,000 markers for an Interior spruce (*Picea glauca* (Moench) Voss x *Picea engelmanni*i Parry ex Engelm.) population, with an N_e ~93. Indeed Howe et al. [17] concluded that a density of 2.5 markers per cM (5,000 SNPs/2,000 cM), should provide effective GS results in populations no larger than N_e ~30. These numbers, for populations with low N_e, are in line with Meuwissen's [9] determination that $10N_eL$ markers should be used as a minimum, which gives us 10×21×20 = 4,200 markers for Douglas-fir. Yet using the same calculation from Meuwissen [9] for Interior spruce, gives us a recommended minimum of 18,600 markers, less than half of that recommended by Grattapaglia and Resende's [12] calculation.

Ma et al. [18] investigated the effect of marker selection on prediction accuracy of GS in soybean (*Glycine max L.*). They tested three methods of marker selection: random sampling, haplotype block analysis, and evenly sampling markers. They found that for plant height, only marginal differences in prediction accuracy were obtained with the three sampling methods. However, for grain yield, the haplotype block analysis out-performed the other two methods by about 4%. This preselection method offers a comprehensive, yet cost-efficient, option for implementing GS by reducing the number of SNPs required to just one SNP per haplotype block plus those not contained within blocks. However, an in-depth understanding of the structure and LD across the genome is required for this. For this reason, we have concentrated only on the random sampling method for Douglas-fir and Interior spruce where current genome assemblies are highly fragmented and not conducive to analyses of genome wide patterns of LD.

The two species studied here are representative of full-sib (Douglas-fir) and half-sib (Interior spruce) populations. Their differing pedigree structure should be reflected in the prediction accuracies we obtain through GS. Previously, using full-sib families, GS prediction accuracy has been shown to be moderate to high in general [19,20], and in Douglas-fir specifically [21]. This is considered to be a result, primarily, due to long range LD arising from the increased levels of relatedness within families. Short range LD is not considered a strong influence in these circumstances. By comparison, using half-sib families has previously resulted in low to moderate GS prediction accuracies in general [22], and in Interior spruce specifically [23,24]. Higher N_e and subsequently lower LD are thought to impede prediction accuracies in studies based on these half-sib populations. Larger N_e leads to more recombination and therefore more diversity within a population. Drift associated with small population size is not a significant factor under these conditions (open-pollinated and highly outcrossing species) and subsequently do not significantly contribute to the build-up of LD. Nonetheless, as can be seen in Fig 1, high GS prediction accuracies are not exclusive to full-sib studies and *vice versa*.

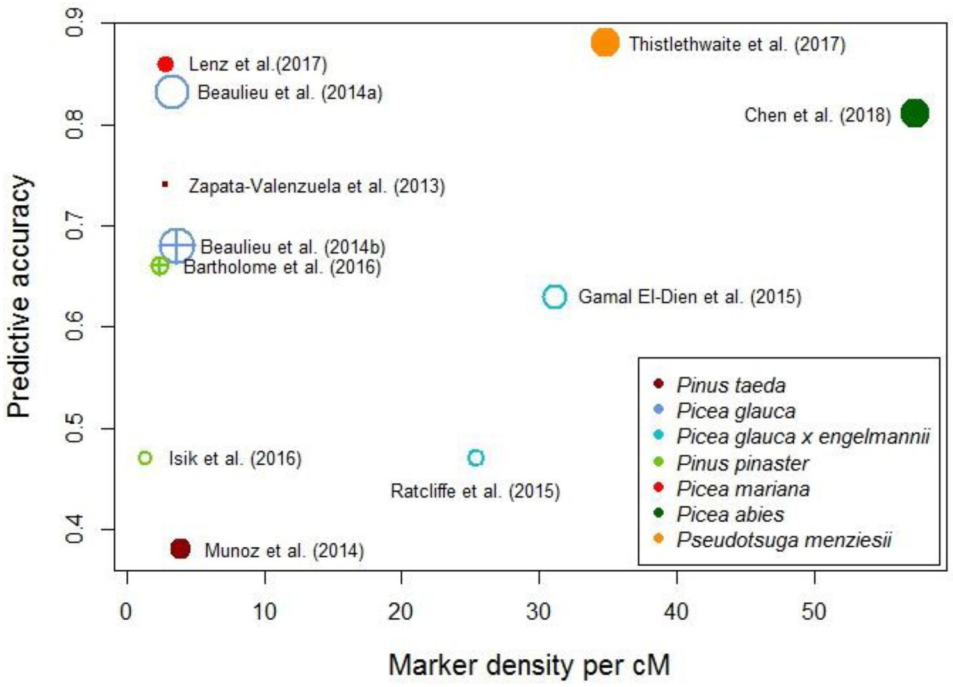

Fig 1. Scatterplot of metadata concerning height prediction accuracy, from various sources of GS studies [19,21–30] in forestry of conifer species. Filled points represent studies using full-sib populations; empty points represent studies using half-sib populations; empty points with crosses represent studies using full and half-sib populations. Point diameter is a function of sample size. Marker density was calculated based on genetic map lengths estimated in the following studies: *Pinus taeda* [31], *Picea glauca* [32], *Picea glauca* x *engelmannii* (based on white spruce data) [32–34], *Pinus pinaster* [35], *Picea mariana* [36], *Picea abies* [37].

https://doi.org/10.1371/journal.pone.0232201.g001

It is, however, becoming seemingly more apparent that perhaps LD is not the main driving force in all GS studies [21,30]. Studies concerning those species with larger genomes may find that relatedness rather than LD is the most important factor influencing prediction accuracies. As de los Campos et al. [31] describe, the success of GS relies upon the similarity of the realised genomic relationships at the marker and QTL levels. The number of independently segregating segments determines the coefficient of variation of these relationships across the genome. In unrelated individuals, this is a product of population-wide LD whereas between family members, this is a product of within family disequilibrium. Here we investigate the effect of marker density on GS prediction accuracy in two conifer species (Douglas-fir and Interior spruce) with differing pedigree structures (full- and half-sib family structure). The main underlying assumption of GS studies is the presence of LD between markers and causal genes, thus the derived predictive models are transferable and can be used for phenotypic prediction of genotypes but non-phenotyped individuals. However, if the obtained predictive accuracy is the result of increased pedigree resolution, then these models should be used with caution, as their predictive accuracy is pedigree dependent.

Results

Marker number effect

For Douglas-fir, the average prediction accuracy for height genomic estimated breeding values (GEBVs) ranged from 0.63 (standard error: ±0.040) to 0.87 (±0.008), and for Interior spruce 0.31 (±0.028) to 0.63 (±0.019), representative of GS testing SNP sets of 200 and 50,000 SNP

number totals, respectively, for each species. Similarly, for Douglas-fir, the average accuracy of wood density GEBVs ranged from 0.62 (±0.023) to 0.83 (±0.011), and for Interior spruce, 0.29 (±0.040) to 0.62 (±0.017), using SNP set totals of 200 and 50,000, respectively. Accordingly, the effect of marker number on the prediction accuracy of multi-site cross-validation data, showed a clear trend of increased predictive accuracy with increasing marker density. However, the magnitude of prediction accuracy gains starts to plateau around a threshold of around 10,000–15,000 markers for both Douglas-fir and Interior spruce (Fig 2). Despite the random selection of markers, there is also little variation in predictive accuracy among SNP sets of the same SNP number total, as indicated by the small error bars in Fig 2. On average, the accuracies for height increased by a factor of 1.03 and 1.08, respectively, when the number of markers was doubled, for Douglas-fir and Interior spruce, respectively. For wood density the accuracies increased on average by a factor of 1.02 and 1.07, respectively, when the number of markers was doubled, for Douglas-fir and Interior spruce, respectively. Fig 2 provides a graphical summary of the results, depicting the average predictive accuracy of 100 replicates per SNP set and 10 random replicates for each SNP set total, as well as their standard errors.

Pedigree and relatedness effect

The full-sib structure related to the Douglas-fir models outperformed those of the half-sib structure of the Interior spruce models for all SNP set totals (Fig 2). The prediction accuracies for the full-sib models (Douglas-fir) were for height and wood density, respectively, on average 1.58 and 1.54 times larger than for the half-sib models (Interior spruce). Prediction accuracy varied more within traits when using half-sibs compared to full-sib pedigrees, as represented by the SNP set total error bars in Fig 2. On average, the standard deviations for half-sib models were 2.59 times larger than those of the full-sib models for height, and 1.75 times larger for wood density.

Discussion

Generally, prediction accuracies for height and wood density GEBVs of both Douglas-fir and Interior spruce, increased with increasing number of SNPs (Fig 2). The data suggests that increasing the number of markers will therefore provide more accurate predictions. However, in light of current genotyping costs and efficiency, it is prudent to use as few markers as possible without loss of significant accuracy. Ideally, this would mean that the minimum number of markers should be the same as the number of independent linkage blocks. This number may vary according to the type of breeding population, and type of genomic data collected. Although the saturation point for our data was around 10,000 markers for both Douglas-fir and Interior spruce using EBVs as the model input, certainly the N_e of the populations need to be considered. Data with a higher N_e (>93) may require more markers to obtain similar prediction accuracies.

The observed GEBV coalesce similar results obtained in previous breeding program studies and generalized simulations [8,22,30,38–42]. The GS prediction accuracies for both height and wood density, parallel those obtained by ABLUP and by GS in previous studies on these species [21,23]. These results from models trained on EBVs led us to conclude that, even at a relatively low density, the markers used were able to capture the genetic relationship as effectively as the pedigree, and in the case of HT at 35 years for Douglas-fir, more effectively. As a result of this finding, future selections may be conducted more efficiently and without the need for a structured pedigree, thereby speeding up the breeding process by eliminating the need to conduct specific crosses [43]. Furthermore, prior studies in Douglas-fir modeling de-regressed EBVs, where parental averages are removed, showed extremely low GS prediction accuracies and largely undetectable LD [21,44].

The role of short-range marker-QTL LD on GS prediction accuracy was imperceptible compared to the effect of relatedness. This is a somewhat anticipated result given that most conifers are relatively undomesticated even within breeding programs, and are known to be

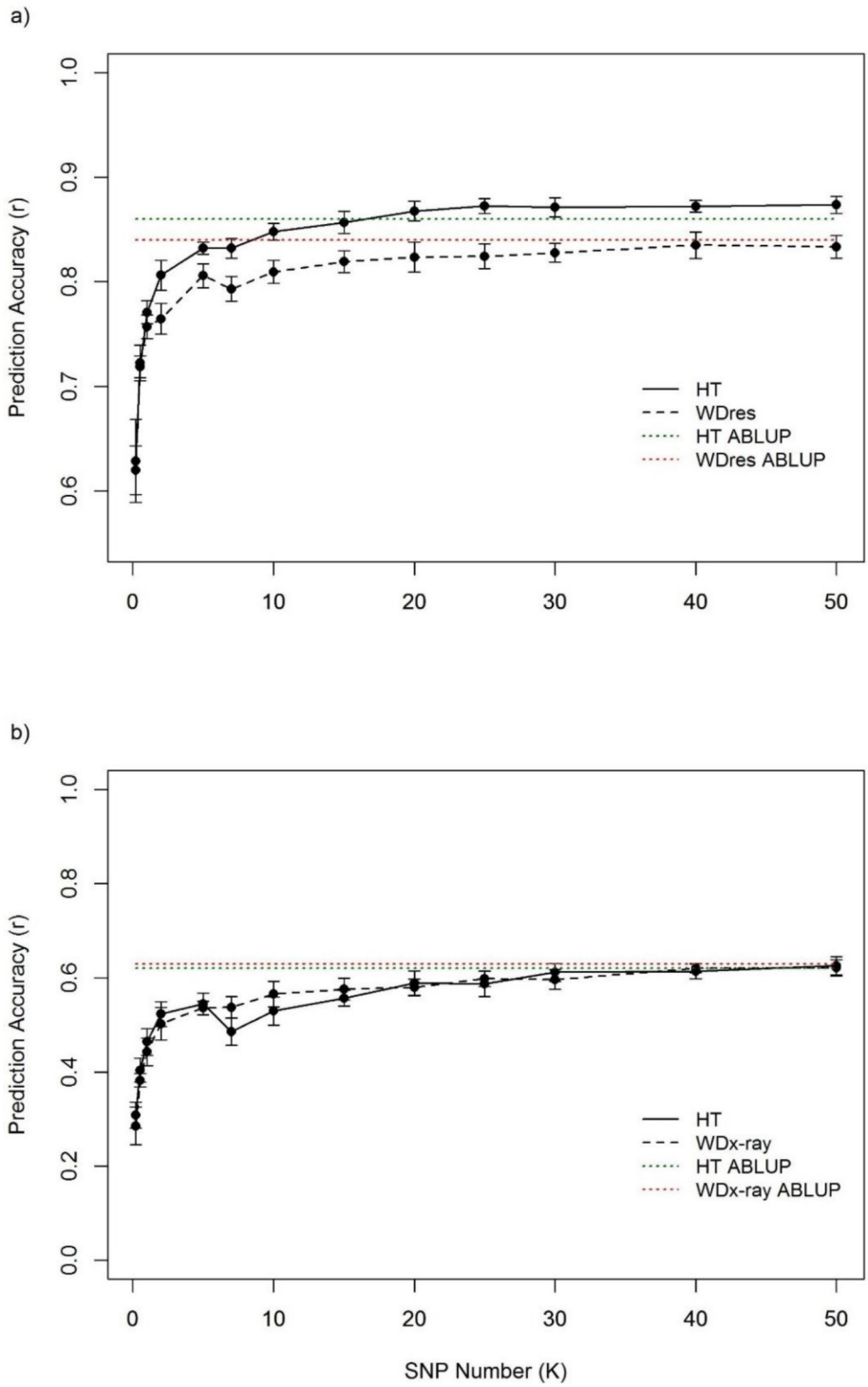

Fig 2. Effect of marker density on RR-BLUP prediction accuracy for multi-site data for height and wood resistance, with ABLUP for comparison, for: a) Douglas-fir (HT at 35 years, WD_{res} at 38 years), and b) Interior spruce (HT at 38 years, $WD_{X\text{-}ray}$ at 38 years).

https://doi.org/10.1371/journal.pone.0232201.g002

highly outbred species with large N_e. These characteristics actively preclude the buildup of population-wide LD [45,46]. Corroborating evidence for this can been seen in Fig 2, where the variance of the prediction accuracies is modest despite SNPs being randomly selected. This leads us to believe that the SNPs are tracking genetic relatedness as opposed to marker-QTL LD. In addition, other studies have drawn similar conclusions that marker selection strategies have little to no effect on prediction accuracy for growth and wood quality traits [19,30,40]. That is to say, that marker sets made up of only those markers with the largest effects show only a minor advantage over using all markers available [19,30]. As reported by Lenz et al. [30], this observation is likely due to those high-effect markers having higher mean allele frequecies, and hence higher information value allowing them to trace genetic relationships efficiently. In addition, Zapata-Valenzuela et al. [40] found no discernable difference in prediction accuracy when using subsets of markers rather than the whole compliment of markers available to them. This was regardless of whether or not those markers were in association with the trait in question. Going further, considering the metadata presented in Fig 1, marker density at the levels currently applied appears to have little to no bearing on GS prediction accuracy. Studies with less dense marker coverage, in some cases over 20 times less dense than the highest, display similarly high prediction accuracies. Since it is unlikely that such few markers could account for LD with most causal variants across such vast genomes, it is an indication that these markers are tracking something else than LD between markers and QTLs (i.e., pedigree).

Even in species, which do not display the complexity of conifer genomes, similar trends due to marker density reduction have been seen. Using two Eucalyptus species and their F_1 hybrids, Tan et al. [41] found no major advantage in using more than 5,000 SNPs compared to using the full data set of 41,304 SNPs available to them. Lorenz et al. [47] noted that in a barley (*Hordeum vulgare* L.) population with high LD, the number of markers used could be reduced to 384, with minor impact on prediction accuracy for disease resistance (note, barley is a selfer). Marker selection via a *k*-means clustering algorithm to sample LD space had only very modest advantages over random sampling. Additionally, in dairy breeding, multiple studies have shown that prediction accuracy can be maintained despite significantly reducing marker density [48,49]. In these cases, it is the effect of low N_e [11] which is driving the prediction accuracy. With low N_e, fewer independent genomic segments arise, decreasing the number of markers required to track these segments [42]. These populations all exhibit closed breeding and subsequently have lower N_e than our sampled populations. Lowering N_e in this way would be disadvantageous when predicting across families.

Complex traits are thought to be largely governed by noncoding variants seemingly affecting gene regulation and expression [50,51]. The number of such variants are thought to be very large, evenly distributed across the genome and have small effect sizes. The heritability of complex traits is thus similarly spread out throughout the entire genome [50]. The upshot of this is the implication that a large proportion of all genes contribute to variation in complex traits. This is at odds with the expectation that trait variants are located within specific and biologically relevant genes and pathways [50]. Yet Tan et al. [41] generally found that SNPs located in intergenic regions provided slightly better prediction accuracies over those located within/near genic regions, or when using all SNPs available. They attributed this to a slower decline of LD in intergenic regions compared to other genome locations, allowing markers in intergenic regions to effectively trace QTLs over longer genomic segments than markers in genome regions with higher rates of LD decay. Similarly Boyle et al. [50], in their summary report, state that SNPs that contribute most to the heritability are often spread widely across the genome and are not closely located to genes with trait-specific functions.

Although genome-wide variation in LD is virtually unknown in conifers, data from other plant species suggest that LD is higher but also more variable in intergenic compared to genic

regions. There are several reasons for this. First, intergenic regions in many species are replete with repetitive elements, mainly various Long Terminal Repeats (LTR) transposable elements [52,53], and such regions are often highly heterochromatic and show reduced cross-over rates. Second, complex structural variation generated by the action of repetitive elements will further limit cross over due to lack of sequence homology in intergenic regions, directing recombination to genomic regions with high gene densities. Both features are well documented in maize, a species that also has a relatively high repetitive genome fraction [52,54]. As noted above, levels of linkage disequilibrium in conifers are poorly studied to date. Small, targeted re-sequencing of exomic regions have found that LD decays relatively rapidly, over a few kilobases at the most [53,55,56]. This is to be expected if most recombination is largely directed towards genic regions. However, since genic regions only constitute at most a percent or so of a typical conifer genome, these rates are likely not representative for most of the genome. As an example, Fu et al. [52] concluded that the repetitive DNA in maize, while constituting the bulk of the genome, likely contributes little if anything to genetic length. Given these observations, where current genome assemblies are highly fragmented and not conducive to analyses of genome wide patterns of LD, it is perhaps not unusual that, despite relatively large marker totals in use here (and an even greater total used in Thistlethwaite et al. [21]) and for Douglas-fir those markers being located in the exonic regions, LD was not successfully traceable.

To further elucidate the role of markers on resolving the pedigree which in turn affected height and wood density predictive accuracies for the two studied species. It should be noted that the Douglas-fir full-sib outperformed the Interior spruce half-sib structure (on average full-sibs were 1.58 and 1.54 times larger than half-sib for height and wood density, respectively) (Fig 2). These results are indicative of the markers ability in resolving higher relationships among the 37 full-sib families (within and cross families due to common parentageas well as hidden inbreeding). On the other hand, while the Interior spruce has fewer families (25 half-sibs), their open-pollinated nature precluded the development of finer relationships as each seed-donor originated from different location, thus distant relationships were not present. This is also clearly demonstrated by the differences in full- and half-sib families N_e values (21 vs. 93 for the Douglas-fir and Interior spruce, respectively).

Conclusions

Although advantageous within a population, reducing marker density may not be the most effective or economical method of carrying out GS, especially in conifers. As mentioned previously we have yet to trace LD with the current array of markers available, only genetic relationships, which is not the intended use of GS. Given the genetic diversity of conifer species, it would perhaps be more prudent to create denser marker arrays that can be used across populations and breeding programs [57]. Increasing the number of markers in such a way could enable us to tease apart the impact of genetic relationship from LD and to investigate multiple traits including unanticipated traits [22]. If this can be achieved, the higher density of markers would offset somewhat the effects of marker-QTL LD decay due to selection and recombination over multiple generations [8,22,57]. Therefore, low density marker arrays would have more impact in more advanced breeding programs [57].

Material and methods

Experimental populations

Predictive models for GS were trained on two progeny testing populations. The first population consist of 38-year-old coastal Douglas-fir (*Pseudotsuga menziesii* Mirb. (Franco)). This population was originally established by the Ministry of Forests, Lands and Natural Resource

Operations of British Columbia, Canada in 1975 and it is made up of 165 full-sib families (54 parents), from which 37 families were selected for sampling from three test sites (**Adams** (Lat. 50 24'42" N, Long. 126 09' 37"W, Elev. 576 mas), **Fleet River** (Lat. 48 39'25" N, Long. 128 05'05" W, Elev. 561 mas), and **Lost Creek** (Lat. 49 22'15" N, Long. 122 14'07" W, Elev. 424 mas)) giving a total of 1,372 trees (N≈500 per site).

The second population consisted of 1,126 38-year-old Interior spruce trees (*Picea glauca* (Moench) Voss x *Picea engelmanni*i Parry ex Engelm.) (N≈375 per site). This progeny test trial was established in 1972/73 by the Ministry of Forests, Lands and Natural Resource Operations of British Columbia Canada and is made up of 181 open-pollinated families, of which the best performing 25 families were selected based on their growth attributes. The trial is located on three sites (**Aleza Lake** (Lat. 54˚ 03′ 15.7″ N, Long. 122˚ 06′ 35.4″ W, Elev. 700 mas), **Prince George Tree Improvement Station (PGTIS**) (Lat. 53˚ 46′ 17.9″ N, Long. 122˚ 43′ 07.6″W, Elev. 610 mas), and **Quesnel** (Lat. 52˚ 59′ 27.2″ N, Long. 122˚ 12′ 30.6″ W, Elev. 915 mas)).

Access to Douglas-fir and Interior spruce progeny test trials was granted by The Ministry of Forests, Lands and Natural Resource Operations of British Columbia, Canada, and all ethics standards have been met.

Phenotyping and genotyping

Mid-rotation height measurements of the sampled trees were recorded: at age 35 for the Douglas-fir population, and at age 38 Interior spruce (HT: in meters). Estimated breeding values (EBVs) for HT were obtained in ASReml 4.0 [58] and used as phenotypes for the genomic prediction analysis. Wood density (WDres) measurements for the Douglas-fir population were taken indirectly, using the average of resistance measurements obtained with a Resistograph® (Instrumenta Mechanik Labor, Germany). Recordings from the Resistograph® were scaled by DBH measurements to obtain wood density indices following El-Kassaby et al. [59]. Wood density in the Interior Spruce population was measured directly in kg/m^3 using X-ray densitometry ($WD_{X\text{-}ray}$), which uses increment cores extracted from the trees.

Genotyping of the Douglas-fir samples, using whole exome capture, was performed in a commercial facility (RAPiD Genomics©, Florida, US), probes were designed based on the available Douglas-fir transcriptome assembly [17]. A total 'pool' of 56,454 SNPs, with <0.40 heterozygosity, was used in this study. For a complete discription of the genotyping process see Thistlethwaite et al. [21] and Neves et al. [60] for the exome capture methodology respectively.

The Interior spruce samples were genotyped via multiplexed, high-throughput Genotyping-by-Sequencing (GBS) following Elshire et al. [61] and Chen et al. [62], on the Illumina HiSeq 2000 at the Cornell University Genomics Core Laboratory (Gamal El-Dien et al. [23]).

Effective population size estimation

The effective population size (N_e) for the Douglas-fir and spruce were estimated using an Excel program developed by Dr. M. Lstiburek (Faculty of Forestry and Wood Sciences, Czech University of Life Sciences Prague, Prague, Czech Republic) that was based on Lindgren et al. [63] status number concept.

Random marker sampling

The effect of the number of markers on predictive accuracy was ascertained by carrying out a random sampling method for choosing markers from the total 'pool' containing 56,454 SNPs for Douglas-fir and 62,190 for Interior spruce. Sets with SNP totals in the range of 200–50,000 were tested and replicated 10 times, randomly sampling SNPs for each repetition. The cross-

validation processes of the RR-BLUP model was then performed using these randomly sampled SNP sets. This analysis was carried out on the height and wood density phenotypes. Assuming a genome length of ~2,000cM for both Douglas-fir [13] and Interior spruce [14–16] (an approximation based on *Picea glauca*, *Pinaceae* data), the average marker densities tested ranged from 0.05–25 markers/cM.

Estimated Breeding Value (EBV) calculation and ABLUP accuracy

EBVs were calculated in ASReml 4.0 [58] via linear mixed model analysis. For the Douglas-fir population the following model was used:

$$y = X\beta + Z_1 a + Z_2 sa + Z_3 s(rep) + Z_4 sf + Z_5 f + e \quad (1)$$

Where y is the phenotypic trait measurement, β is a vector of fixed effects (including mean and site effects), a is a vector of individual random additive genetic effects ~ $N(0, A\sigma_a^2)$, sa is a site x additive genetic interaction ~ $N(0, I\sigma_{sa}^2)$, $s(rep)$ is a vector of the block effect within site ~ $N(0, I\sigma_{s(rep)}^2)$, sf is a random effect site x family interaction ~ $N(0, I\sigma_{sf}^2)$, f is the effect of family ~ $N(0, I\sigma_f^2)$, and e is the random residual effect ~ $N(0, I\sigma_e^2)$, and X, $Z_{1\text{-}5}$ are incidence matrices assigning fixed and random effects to each observation. I is the identity matrix and A the average numerator relationship matrix.

Narrow-sense heritability was calculated as $h^2 = \sigma_a^2 / (\sigma_a^2 + \sigma_{sa}^2 + \sigma_{sf}^2 + \sigma_f^2 + \sigma_e^2)$, where σ_a^2, σ_{sa}^2, σ_{sf}^2, σ_f^2 and σ_e^2 are the variances of additive genetic, site x additive genetic, site x family, family, and residual effects, respectively.

For the Interior Spruce population, a similar mixed model was used, without family effect:

$$y = X\beta + Z_1 a + Z_2 sa + Z_3 s(rep) + e \quad (2)$$

where y is the phenotypic measurement of the analyzed trait, β is a vector of fixed effect (i.e., the overall mean and the site effect), a is a vector of random additive effects ~ $N(0, A\sigma^2_a)$, sa is a site x additive genetic interaction ~ $N(0, I\sigma^2_{sa})$, $s(rep)$ is a vector of the block effect within site ~ $N(0, I\sigma_{s(rep)}^2)$, e is a the random residual effect ~ $N(0, I\sigma^2_e)$, and X, $Z_{1\text{-}3}$ are incidence matrices assigning fixed and random effects to each observation. I is the identity matrix and A the average numerator relationship matrix.

Narrow-sense heritability was calculated as $h^2 = \sigma_a^2 / (\sigma_a^2 + \sigma_{sa}^2 + \sigma_e^2)$, where σ_a^2, σ_{sa}^2, and σ_e^2 are the variances of additive genetic, site x additive genetic, and residual effects, respectively.

ABLUP cross-validation for both species was performed in ASReml R v4.0 [58]. Ten fold cross-validation was performed, using randomly sampled individuals from all 3 sites (Adams, Fleet River and Lost Creek for Douglas-fir; Aleza Lake, PGTIS and Quesnel for Interior spruce) to construct the model, and the remainder to compose the validation set. Prediction accuracy for ABLUP was calculated as the correlation between the EBVs from the validation sets and their original EBVs calculated from Eqs 1 and 2, for Douglas-fir and Interior spruce respectively. ABLUP prediction accuracy was compared to GS prediction accuracy for all SNP set totals.

Genomic selection and cross-validation

The GS method used in the analysis was ridge regression (RR-BLUP) [64], and was implemented using the R package 'bigRR' [65]. The genomic predicted EBVs (GEBVs) for height, were calculated as the sum of all marker effects within each individual tree. Marker effects were estimated using the following mixed model, from Henderson [66]:

$$y_{EBV} = 1\mu + Zg + e \quad (3)$$

where $\boldsymbol{y_{EBV}}$ is the vector of n tree EBV records for height, $\mathbf{1}$ is a vector of 1, $\boldsymbol{\mu}$ is the intercept, $\boldsymbol{g}$ is the vector of random marker effects, $\boldsymbol{Z}$ is the design matrix for the random marker effects, and $\boldsymbol{e}$ is the residual random effects vector. In the RR-BLUP procedure, the residuals and marker effects are presumed to follow normal distributions with constant variance, i.e. $e \sim \mathrm{N}(0, \boldsymbol{I}\sigma_e^2)$ and $g \sim \mathrm{N}(0, \boldsymbol{I}\sigma_g^2)$, where $\boldsymbol{I}$ is an identity matrix. Marker effect solutions are calculated according to the following equation:

$$\hat{\boldsymbol{g}} = (\boldsymbol{Z}'\boldsymbol{Z} + \lambda \boldsymbol{I})^{-1}\boldsymbol{Z}\prime\boldsymbol{y} \tag{4}$$

where $\lambda = \sigma_e^2/\sigma_g^2$ is the ridge penalization parameter. Marker effects are assumed to be distributed equally, and as such all are uniformly shrunk towards zero.

Predictive accuracy was used to estimate the performance of this GS method. Predictive accuracy was determined as the mean of the replications of the Pearson product-moment correlation between estimated breeding values (EBVs) of the validation set and their genomic estimated breeding values (GEBVs) or r(EBV,GEBV). Validation was applied as a replicated randomly assigned 10-fold cross validation repeated 10 times in which 9/10 folds were used to train the model, the other fold constituting the validation population. Information from the 3 sites were pooled.

Acknowledgments

We thank British Columbia Ministry of Forests, Lands and Natural Resource Operations, Victoria, BC for data and trials access.

Author Contributions

Conceptualization: Frances R. Thistlethwaite, Blaise Ratcliffe, Jaroslav Klápště, Ilga Porth, Charles Chen, Michael U. Stoehr, Pär K. Ingvarsson, Yousry A. El-Kassaby.

Data curation: Omnia Gamal El-Dien, Blaise Ratcliffe.

Formal analysis: Frances R. Thistlethwaite, Omnia Gamal El-Dien.

Methodology: Yousry A. El-Kassaby.

Supervision: Yousry A. El-Kassaby.

Writing – original draft: Frances R. Thistlethwaite, Omnia Gamal El-Dien.

Writing – review & editing: Blaise Ratcliffe, Jaroslav Klápště, Ilga Porth, Charles Chen, Michael U. Stoehr, Pär K. Ingvarsson, Yousry A. El-Kassaby.

References

1. Grattapaglia D. Breeding forest trees by genomic selection: Current progress and the way forward. In: Genomics of Plant Genetic Resources (Tuberosa R, Graner A, Frison E, eds). Dordrecht: Springer Netherlands; 2014:651–682.
2. Heffner EL, Lorenz AJ, Jannink J-L, Sorrells ME. Plant breeding with genomic selection: Gain per unit time and cost. Crop Sci. 2010; 50:1681.
3. El-Kassaby YA. Associations between allozyme genotypes and quantitative traits in Douglas-fir [*Pseudotsuga menziesii* (Mirb.) Franco]. Genetics. 1982; 101:103–115. PMID: 17246076
4. Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001; 157:1819–1829. PMID: 11290733
5. Jannink J-L, Lorenz AJ, Iwata H. Genomic selection in plant breeding: From theory to practice. Brief Funct Genomics. 2010; 9:166–177. https://doi.org/10.1093/bfgp/elq001 PMID: 20156985
6. Lin Z, Hayes BJ, Daetwyler HD. Genomic selection in crops, trees and forages: A review. Crop Pasture Sci. 2014; 65:1177.

7. Falconer DS, Mackay TFC. Introduction to quantitative genetics. Harlow: Pearson, Prentice Hall; 2009. https://doi.org/10.1038/ng.352

8. Solberg TR, Sonesson AK, Woolliams JA, Meuwissen THE. Genomic selection using different marker types and densities. J Animal Sci. 2008; 86:2447–2454.

9. Meuwissen TH. Accuracy of breeding values of "unrelated" individuals predicted by dense SNP genotyping. Genet Sel Evol. 2009; 41:35. https://doi.org/10.1186/1297-9686-41-35 PMID: 19519896

10. Kumar S, Chagné D, Bink MCAM, Volz RK, Whitworth C, Carlisle C. Genomic selection for fruit quality traits in apple (*Malus*×*domestica* Borkh.). PLoS ONE. 2012; 7:e36674. https://doi.org/10.1371/journal.pone.0036674 PMID: 22574211

11. Hayes BJ, Visscher PM, Goddard ME. Increased accuracy of artificial selection by using the realized relationship matrix. Genet Res. 2009; 91:47.

12. Grattapaglia D, Resende MDV. Genomic selection in forest tree breeding. Tree Genet Genomes. 2011; 7:241–255.

13. Krutovsky KV, Troggio M, Brown GR, Jermstad KD, Neale DB. Comparative mapping in the *Pinaceae*. Genetics. 2004; 168:447. https://doi.org/10.1534/genetics.104.028381 PMID: 15454556

14. Pavy N, Pelgas B, Beauseigle S, Blais S, Gagnon F, Gosselin I, et al. Enhancing genetic mapping of complex genomes through the design of highly-multiplexed SNP arrays: Application to the large and unsequenced genomes of white spruce and black spruce. BMC Genomics. 2008; 9:21. https://doi.org/10.1186/1471-2164-9-21 PMID: 18205909

15. Pavy N, Lamothe M, Pelgas B, Gagnon F, Birol I, Bohlmann J, et al. A high-resolution reference genetic map positioning 8.8 K genes for the conifer white spruce: structural genomics implications and correspondence with physical distance. Plant J. 2017; 90:189–203. https://doi.org/10.1111/tpj.13478 PMID: 28090692

16. Pelgas B, Bousquet J, Meirmans PG, Ritland K, Isabel N. QTL mapping in white spruce: Gene maps and genomic regions underlying adaptive traits across pedigrees, years and environments. BMC Genomics. 2011; 12:145. https://doi.org/10.1186/1471-2164-12-145 PMID: 21392393

17. Howe GT, Yu J, Knaus B, Cronn R, Kolpak S, Dolan P, et al. A SNP resource for Douglas-fir: de novo transcriptome assembly and SNP detection and validation. BMC Genomics. 2013; 14:137. https://doi.org/10.1186/1471-2164-14-137 PMID: 23445355

18. Ma Y, Reif JC, Jiang Y, Wen Z, Wang D, Liu Z, et al. Potential of marker selection to increase prediction accuracy of genomic selection in soybean (*Glycine max* L.). Mol Breed. 2016; 36:113. https://doi.org/10.1007/s11032-016-0504-9 PMID: 27524935

19. Beaulieu J, Doerksen TK, MacKay J, Rainville A, Bousquet J. Genomic selection accuracies within and between environments and small breeding groups in white spruce. BMC Genomics. 2014; 15:1048. https://doi.org/10.1186/1471-2164-15-1048 PMID: 25442968

20. Resende MFR Jr, Muñoz P, Acosta JJ, Peter GF, Davis JM, Grattapaglia D et al. Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. New Phytol. 2012; 193:617–624. https://doi.org/10.1111/j.1469-8137.2011.03895.x PMID: 21973055

21. Thistlethwaite FR, Ratcliffe B, Klápště J, Porth I, Chen C, Stoehr MU, et al. Genomic prediction accuracies in space and time for height and wood density of Douglas-fir using exome capture as the genotyping platform. BMC Genomics. 2017; 1:930.

22. Beaulieu J, Doerksen T, Clément S, MacKay J, Bousquet J. Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. Heredity. 2014; 113:343–352. https://doi.org/10.1038/hdy.2014.36 PMID: 24781808

23. Gamal El-Dien O, Ratcliffe B, Klápště J, Chen C, Porth I, El-Kassaby YA. Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. BMC Genomics. 2015; 16:370. https://doi.org/10.1186/s12864-015-1597-y PMID: 25956247

24. Ratcliffe B, Gamal El-Dien O, Klápště J, Porth I, Chen C, Jaquish B, et al. A comparison of genomic selection models across time in interior spruce (*Picea engelmannii* × *glauca*) using unordered SNP imputation methods. Heredity. 2015; 115:547–555. https://doi.org/10.1038/hdy.2015.57 PMID: 26126540

25. Bartholomé J, Van Heerwaarden J, Isik F, Boury C, Vidal M, Plomion C, et al. Performance of genomic prediction within and across generations in maritime pine. BMC Genomics. 2016; 17:604. https://doi.org/10.1186/s12864-016-2879-8 PMID: 27515254

26. Chen Z-Q, Baison J, Pan J, Karlsson B, Andersson B, Westin J, et al. Accuracy of genomic selection for growth and wood quality traits in two control-pollinated progeny trials using exome capture as the genotyping platform in Norway spruce. BMC Genomics. 2018; 19.

27. Isik F, Bartholomé J, Farjat A, Chancerel E, Raffin A, Sanchez L, et al. Genomic selection in maritime pine. Plant Science. 2016; 242:108–119. https://doi.org/10.1016/j.plantsci.2015.08.006 PMID: 26566829

28. Munoz PR, Resende MFR, Huber DA, Quesada T, Resende MDV, Neale DB, et al. Genomic relationship matrix for correcting pedigree errors in breeding populations: Impact on genetic parameters and genomic selection accuracy. Crop Science. 2014; 54:1115.

29. Zapata-Valenzuela J, Whetten RW, Neale D, McKeand S, Isik F. Genomic estimated breeding values using genomic relationship matrices in a cloned population of loblolly pine. Genes|Genomes|Genetics. 2013; 3:909–916. https://doi.org/10.1534/g3.113.005975 PMID: 23585458

30. Lenz PRN, Beaulieu J, Mansfield SD, Clément S, Desponts M, Bousquet J. Factors affecting the accuracy of genomic selection for growth and wood quality traits in an advanced-breeding population of black spruce (*Picea mariana*). BMC Genomics 2017; 18:335. https://doi.org/10.1186/s12864-017-3715-5 PMID: 28454519

31. de los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D. Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet. 2013; 9:e1003608. https://doi.org/10.1371/journal.pgen.1003608 PMID: 23874214

32. Eckert AJ, van Heerwaarden J, Wegrzyn JL, Nelson CD, Ross-Ibarra J, Gonzalez-Martinez SC, et al. Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). Genetics. 2010; 185:969–982. https://doi.org/10.1534/genetics.110.115543 PMID: 20439779

33. Pavy N, Lamothe M, Pelgas B, Gagnon F, Birol I, Bohlmann J, et al. A high-resolution reference genetic map positioning 8.8 K genes for the conifer white spruce: structural genomics implications and correspondence with physical distance. The Plant Journal. 2017; 90:189–203. https://doi.org/10.1111/tpj.13478 PMID: 28090692

34. Pavy N, Pelgas B, Beauseigle S, Blais S, Gagnon F, Gosselin I, et al. Enhancing genetic mapping of complex genomes through the design of highly-multiplexed SNP arrays: application to the large and unsequenced genomes of white spruce and black spruce. BMC Genomics. 2008; 9:21. https://doi.org/10.1186/1471-2164-9-21 PMID: 18205909

35. Pelgas B, Bousquet J, Meirmans PG, Ritland K, Isabel N. QTL mapping in white spruce: gene maps and genomic regions underlying adaptive traits across pedigrees, years and environments. BMC Genomics. 2011; 12.

36. Ritter E, Aragones A, Markussen T, Achere V, Espinel S, Fladung M, et al. Towards construction of an ultra high density linkage map for *Pinus pinaster*. Ann For Sci. 2002; 59:637–643.

37. Kang B-Y, Mann IK, Major JE, Rajora OP. Near-saturated and complete genetic linkage map of black spruce (*Picea mariana*). BMC Genomics. 2010; 11:515. https://doi.org/10.1186/1471-2164-11-515 PMID: 20868486

38. Asoro FG, Newell MA, Beavis WD, Scott MP, Jannink J-L. Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. Plant Genome. 2011; 4:132.

39. Sonesson AK, Meuwissen TH. Testing strategies for genomic selection in aquaculture breeding programs. Genet Sel Evol. 2009; 41:37. https://doi.org/10.1186/1297-9686-41-37 PMID: 19566932

40. Zapata-Valenzuela J, Isik F, Maltecca C, Wegrzyn J, Neale D, McKeand S, et al. SNP markers trace familial linkages in a cloned population of *Pinus taeda*—prospects for genomic selection. Tree Genet Genomes. 2012; 8:1307–1318.

41. Tan B, Grattapaglia D, Martins GS, Ferreira KZ, Sundberg B, Ingvarsson PK. Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F1 hybrids. BMC Plant Biol. 2017; 17:110. https://doi.org/10.1186/s12870-017-1059-6 PMID: 28662679

42. Wang Q, Yu Y, Yuan J, Zhang X, Huang H, Li F, et al. Effects of marker density and population structure on the genomic prediction accuracy for growth trait in Pacific white shrimp *Litopenaeus vannamei*. BMC Genetics. 2017; 18:45. https://doi.org/10.1186/s12863-017-0507-5 PMID: 28514941

43. El-Kassaby YA, Lstibůrek M. Breeding without breeding. Genet Res. 2009; 91:111.

44. Thistlethwaite FR, Ratcliffe B, Klápště J, Porth I, Chen C, Stoehr MU, et al. Genomic selection of juvenile height across a single-generational gap in Douglas-fir. Heredity. 2019; https://doi.org/10.1038/s41437-018-0172-0 PMID: 30631145

45. Neale DB, Savolainen O. Association genetics of complex traits in conifers. Trends Plant Sci. 2004; 9:325–330. https://doi.org/10.1016/j.tplants.2004.05.006 PMID: 15231277

46. Isik F. Genomic selection in forest tree breeding: The concept and an outlook to the future. New For. 2014; 45:379–401.

47. Lorenz AJ, Smith KP, Jannink J-L. Potential and optimization of genomic selection for fusarium head blight resistance in six-row barley. Crop Sci. 2012; 52:1609.

48. Su G, Brøndum RF, Ma P, Guldbrandtsen B, Aamand GP, Lund MS. Comparison of genomic predictions using medium-density (~54,000) and high-density (~777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. J Dairy Sc. 2012; 95:4657–4665.

49. Zhang Z, Ding X, Liu J, Zhang Q, de Koning D-J. Accuracy of genomic prediction using low-density marker panels. J Dairy Sci. 2011; 94:3642–3650. https://doi.org/10.3168/jds.2010-3917 PMID: 21700054

50. Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: From polygenic to omnigenic. Cell. 2017; 169:1177–1186. https://doi.org/10.1016/j.cell.2017.05.038 PMID: 28622505

51. Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am J Hum Genet. 2014; 94:559–573. https://doi.org/10.1016/j.ajhg.2014.03.004 PMID: 24702953

52. Fu H, Zheng Z, Dooner HK. Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. Proc Nat Acad Sci. 2002; 99:1082–1087 https://doi.org/10.1073/pnas.022635499 PMID: 11792865

53. Larsson H, Källman T, Gyllenstrand N, Lascoux M. Distribution of long-range linkage disequilibrium and Tajima's D values in Scandinavian populations of Norway spruce (*Picea abies*). Genes|Genomes|Genetics. 2013; 3:795–806. https://doi.org/10.1534/g3.112.005462 PMID: 23550126

54. Dooner HK, He L. Maize genome structure variation: Interplay between retrotransposon polymorphisms and genic recombination. Plant Cell. 2008; 20:249–258. https://doi.org/10.1105/tpc.107.057596 PMID: 18296625

55. Heuertz M, De Paoli E, Kallman T, Larsson H, Jurman I, Morgante M, et al. Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [*Picea abies* (L.) Karst]. Genetics. 2006; 174:2095–2105. https://doi.org/10.1534/genetics.106.065102 PMID: 17057229

56. Pyhajarvi T, Garcia-Gil MR, Knurr T, Mikkonen M, Wachowiak W, Savolainen O. Demographic history has influenced nucleotide diversity in European *Pinus sylvestris* populations. Genetics. 2007; 177:1713–1724. https://doi.org/10.1534/genetics.107.077099 PMID: 18039881

57. Grattapaglia D. Status and perspectives of genomic selection in forest tree breeding. In Genomic Selection for Crop Improvement, (Varshney RK, Roorkiwal M, Sorrells ME., eds.) Cham: Springer International Publishing. 2017:199–249.

58. Gilmour AR, Gogel BJ, Cullis BR, Thompson R. ASReml User Guide Release 3.0; 2009.

59. El-Kassaby YA, Mansfield S, Isik F, Stoehr M. In situ wood quality assessment in Douglas-fir. Tree Genet Genomes. 2011; 7:553–561.

60. Neves LG, Davis JM, Barbazuk WB, Kirst M. Whole-exome targeted sequencing of the uncharacterized pine genome. Plant J. 2013; 75;146–156. https://doi.org/10.1111/tpj.12193 PMID: 23551702

61. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-Sequencing (GBS) approach for high diversity species. PLoS ONE. 2011; 6:e19379. https://doi.org/10.1371/journal.pone.0019379 PMID: 21573248

62. Chen C, Mitchell SE, Elshire RJ, Buckler ES, El-Kassaby YA. Mining conifers' mega-genome using rapid and efficient multiplexed high-throughput genotyping-by-sequencing (GBS) SNP discovery platform. Tree Gene Genomes. 2013; 9:1537–1544.

63. Lindgren D, Gea LD, Jefferson PA. Status number for measuring genetic diversity. For Genet. 1997; 4:69–76.

64. Whittaker JC, Thompson R, Denham MC. Marker-assisted selection using ridge regression. Genet Res. 2000; 75:249–252. https://doi.org/10.1017/s0016672399004462 PMID: 10816982

65. Shen X, Alam M, Ronnegard L. Package "bigRR": Generalized Ridge Regression (with special advantage for p >> n cases); 2014.

66. Henderson CR. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics. 1976; 32:69.